1,,
Received: from mail.think.com by quake.think.com (4.1/SMI-4.0)
id AA12559; Tue, 6 Oct 92 13:23:31 PDT
Return-Path: <ddern@world.std.com>
Received: from Think.COM by mail.think.com; Tue, 6 Oct 92 16:27:06 -0400
Received: from world.std.com by Early-Bird.Think.COM; Tue, 6 Oct 92 16:26:38 EDT
Received: by world.std.com (5.65c/Spike-2.0)
id AA17810; Tue, 6 Oct 1992 16:26:08 -0400
Date: Tue, 6 Oct 1992 16:26:08 -0400
From: ddern@world.std.com (Daniel P Dern)
Message-Id: <199210062026.AA17810@world.std.com>
To: brewster@Think.COM
Subject: For Review: WAIS chapter for my Internet book

*** EOOH ***
Date: Tue, 6 Oct 1992 16:26:08 -0400
From: ddern@world.std.com (Daniel P Dern)
To: brewster@Think.COM
Subject: For Review: WAIS chapter for my Internet book

Really good!

Daniel P. Dern
DERN ASSOCIATES
P.O. Box 309
Newton Centre, MA 02159

High-Tech Writing Services
(617) 969-7947
FAX: (617) 969-7949
MCIMail: dandern
Internet: ddern@world.std.com

Dern's Reviewer Guidelines:

Brewster,

Dern's Reviewer Guidelines:

DEAR CHAPTER REVIEWER,

Thanks for agreeing to act as a reviewer for drafts of one or

more chapters of my book, The Internet Guide for New Users, before I submit it to McGraw-Hill. (They'll be sending it out to their own set of reviewers, but I want to make sure what I write has been seen by 'net experts' such as yourself.)

Here's some context and requested procedures I'm looking for in having you do this reviewing:

DON'T REDISTRIBUTE:

Because McGraw-Hill has purchased exclusive first rights (and many other exclusive and non-exclusive rights) to this document, I ask that you not casually share the draft text I sent you with other people.

Specially, please don't redistribute the on-line text without checking with me first -- and PLEASE don't post it to newsgroups or mailing lists. Thanks for your cooperation.

If there is someone else you feel should see/review the information, let me know so I can contact them.

(Some of the info is making it into articles I'm selling before the book's publication, and I do intent to put some information up on the net as free/sharetext postings. McGraw Hill has some on-line electronic publishing mechanism, also.)

WHAT KIND OF REVIEWING I'M LOOKING FOR:

In terms of feedback, my greatest interest in accuracy and relevance.

1) If you are quoted anywhere, please verify your quote(s) -- and feel free to change/revise/remove/add to them, plus verify that I've got your "appositive" (so-and-so <netaddress if you want> is a TITLE-TITLE at ORG, CITY, STATE, who <other one notable fact>). This is your 'free commercial' so I want it to be what you most want it to say.

2) I'm looking to you to help catch errors of fact, omissions, and tips/gotchas/shortcuts, as well as flag any statements whose political or cultural content should be tweaked.

3) If you find there are additional comments that belong in the text that merit coming from you, as a quote, create away! (I don't promise to add it, but you're welcome to suggest away.) The information in this book derives from the work of many people; as I hope is evident, I'm doing my best to give credit where do and cite experts.

4) If you feel there are other people or documents that should be referenced or cited, let me know.

5) I also welcome your thoughts on organization, style, tone, grammer, spelling and attitude (mine).

TIMING:

If possible, I would appreciate feedback within a week. My deadline is fast approaching; I can use as much time as possible to apply your feedback. (This bookwriting biz is harder than it looks!)

FEEDBACK FORMAT:

Please use some format that lets me see both my original and your changes; otherwise, I'll go crazy staring at multiple versions and merging fixes. Feel free to also include the reasons flagged as FYI.

If possible, give the correct info as opposed to simply telling me that I'm wrong.

E.g.

Using FTP, you can grab entire directory trees with a single command

YOU>
Not true! With several commands, e.g. tar and then ftp

CONTEXT: Goal of Book

Here's some overview info about the book, as context for your reviewing thought processes:

Audience: The book is intended for novice new users, who may never have used a computer, network or Unix before. (And those who have, of course.) It assumes most readers have specific reasons to get on and use the net, although possibly also interested in the Internet R&R personal/fun stuff.

Contents: I'm including information on how to join/get an account (including locate an appropriate Internet Account Provider, such as NetCom, PANIX, Well or World, or service such as PSILink), including "OuterNet" e-mail/Usenet only options like the "nixpubs." Including "bootstrap" tips and how to get info when you're not yet on the net, or don't even have e-mail... or even a computer/modem yet.

It will include "survival Unix," extensive explanations of FTP, telnet, getting archives, e-mail, netiquette, etc... the usual suspects, plus some other things I feel belong.

It's NOT intended as a "The Compleat Internet Guide." I don't plan to include lists of providers, resources, etc., which will be out of date before the ink dries. I do plan to give meta-references to stable sources of these lists. It won't tell you how to plug your organization in, or how to be an Internet sysadmin, or program, etc. No talk about alt.hamsters.duct.tape or the like. They'll find out about that stuff soon enough.

I'm trying to give a sense of Internet terminology, culture, community, and history, without requiring the reader to already "speak Unix". I also am giving non-Internet context, e.g., how Archie and WAIS can be used in non-Internet situations.

OTHER: REVIEWS, CONTRIBUTIONS, EXCERPTS & PRE-ORDERS

Lastly, a quick note about related topics. If you are interested in getting a review copy of the book when available, let me know. If you think there is some short item by someone else that belongs as an appendix, tell me and I'll work on reprint permissions, rights and payment. I can pre-sell chapters to magazines, etc., within reason -- let me know if you think some editor might be interested. And my contract includes some nifty provision for bulk and custom sales -- I'm ready to start talking

about special arrangements, etc.

Thanks for your patience in reading this.

Daniel P. Dern

Daniel P. Dern
P.O. Box 309
Newton Centre, MA 02159
(617) 969-7947
MCI Mail: dandern
Internet: ddern@world.std.com
FAX: (617) 969-7949

Internet Guide for New Users
Chapter: WAIS
DRAFT - DO NOT REDISTRIBUTE
October 1992
5,000+ words

by Daniel P. Dern

```
*******************************
 *                           *
* DRAFT * DRAFT * DRAFT *
 *                           *
*  NOT FOR REDISTRIBUTION!  *
 *                           *
   * (Please see attached or    *
   * 'nearby' "Dern's Reviewer  *
  * Guidelines")              *
 *                           *
*******************************
```


[For: Internet Navigators & Front Ends section]

WAIS: Finding Needles In Network Haystacks

WAIS (pronounced "ways"), the Wide-Area Information Server system, is one of the solutions to a problem common to any large pool of on-line information, namely, finding something, particularly when you're not sure where it is or exactly what you're looking for -- which is about as precise a description of the distribution of information across the Internet as you can get.

The good news with on-line information is that you can use the computer that's storing it as a tool to look through it -- something my file cabinet won't do. Equally nice, if you have the right tool, and know how to work it, you can crank up the command, and let the computer do the work. (I wish I had a magic spell that would do this for all the paper on my desk. Don't you?)

smooth out this sentence

On small local file systems, where you've only got a few tens of megabytes, Unix's grep command, Mac utilities like ON Location and DOS terminate-and-stay-resident (TSR) file-browser commands can be adequate. On my 286-based PC, for example, QuickSoft's PC-Browse seems to be able to browse through a megabyte of files in a few seconds. On my account on The World, which is a Sun Microsystems Unix system, a _grep_ command searches through my text files with similar rapidity.

And for larger amounts of well-structured data, database management systems like Paradox, Oracle and Progress let users search through millions of records -- but only if you're comfortable with the query language and Boolean logic operators. And if the information's been put into the appropriate database format.

Similarly, on-line database services and catalogs like Dialog and the Research Library Group's RLIN Research Library Information Network let you search -- but it takes expertise, and someone having packaged up and formatted the information properly.

On the Internet, it's even more complicated. (What isn't, you say?) The information you want may in your personal files. Or in directories owned by your organization, e.g., five years' worth of your projects reports and memos, or Human Resources Policies and Procedures manual. Or elsewhere on the Internet (that is, on computers at some other organization and site) in

the archives of directories maintained by one, even several other people or organizations, such as three years of research notes, ten years of RISKS e-mail digests, or the record descriptions from the Archie database of Internet-accessible programs.

Or the information you want may be in a combination of personal, organizational and other locations' files.

What's more, the computers involved may be different types of systems, created by different programs, and perhaps of different data types and formats. This includes the format used by Usenet News postings, e-mail messages, and others peculiar to the Internet.

The problem of finding a needle in a haystack has suddenly become the problem of also knowing where all the haystacks are, and determining which one or ones you want to search through.

To do this, you need a tool that can munch its way across networks and into disparate data formats.

Or you can use WAIS.

WAIS: Developed for Super-Searching

Begun in January 1990, the WAIS project was begun by developers at supercomputer vendor Thinking Machines Corporation (TMC) (Menlo Park, CA), working with Apple Computer and Dow Jones/News
Retrieval on corporate test subject KPMG Peat Marwick, under the leadership of Brewster Kahle (brewster@think.com).

According to Kahle, "Our goal was to to see if we were able to deliver the computing system we'd been promised for two decades, namely, to bring the library to end users' desks, and make it possible for people to access large quantities of text in a user-friendly manner."

"You can view WAIS as a dynamic hypertext system, that creates hypertext links for you," says Kahle. "Or as a corporate memory that allows you to access personal and corporate information and wide area information through one common interface -- something that's never been done before. And you can use WAIS as a huge

world-wide 'find file' system. The information you want is probably out there; WAIS is a way to find it, a way to navigate across databases and the computers that house them, and search through them."

WAIS is a network-based client-server application for full-text retrieval.

that runs on your PC, mac, or unix box

You use the client program to create queries that select which ~~datasets~~ collection (?) to search, enter commands and keywords, etc.

The server handles indexing of datasets, and searching/retrieval in response to queries by users. Indexing may be redone periodically as information is added, changed or deleted.

a list of "headlines"

The server you have specified applies these commands to the designated datasets, and returns ~~its results~~ to your client. where the best matches are listed first. Initially, you see a list of what documents have matched, which you can then select, which causes the document to be displayed at your client. Or you can refine the search, narrowing it down, or causing other documents to be added. Or you can start a brand new search, on the same or new datasets.

WAIS clients and servers communicate using the international standard (Z39.50) protocol, which specifies the format in which WAIS clients send queries to the servers, and in which servers return results and retrieved documents. Any client software that can "speak" Z39.50 can send queries to WAIS servers; similarly, WAIS clients can query any server that "understands" Z39.50, ~~even if they aren't WAIS servers~~.

[IS THE ABOVE CLAIM TRUE? DPD] Not quite. Z39.50 is huge and strange, different people can be "compliant" but not interoperable.

As the query language, WAIS uses natural language (e.g., English, French).

"When we went to Peat Marwick," recalls Kahle, "their CIO said 'No algebra' -- a beautiful way of putting it. 'These partners are not going to learn another query language.' We said, "no problem!" All they do is say what they're interested in. E.g., say contracts about IBM and Motorola. It returns a list, you indicate what you like, and 'find me more like that one.'"

originally pioneered in the 60's + 70's and commercialized in the ...

As its search technique, WAIS employes a method called 'relevance feedback,' ~~previously tried~~ in Dow Jones' DowQuest system. Once you give one or more keywords, WAIS searches through the specified datasets, identifies which documents contain them, and gives you a list, ranking these documents based on how many of the keywords they contains and on other rules.

Relevance feedback means you can indicate which of these search matches you like, and can command WAIS to "find me more like these." The matched documents thus become the specifiers for the next search. "The most relevant documents, regardless of size, can be sent back to the server in their entirety to further refine your search," says Kahle.

"Using this technique, we've gotten end users to locate what they want from within gigabytes of information," states Kahle. "Boolean queries can't provide this; they're only useful for the trained and tolerant, and generally work on fairly small full text databases."

The Internet abounds with large databases. The Archie database of files available from public-access archive sites had over 2.1 million entries by mid-1992. Many of the Internet mailing lists and Usenet News groups have been around for ten or more years, some with dozens of entires per day. Library catalogues and user directory databases likewise run into the million-or-more record sizes. For the high end of the type of information WAIS is intended to handle, the information bases being searched can be large enough to demand to speed of the hundreds to thousands of processors in a TMC Connection Machine. *

FOOTNOTE

This should come as no surprise -- why else would TMC have decided this was an interesting project to pursue?

Thinking Machines' Connection Machines are massively-parallel processing computers. "Parallel-processing" means that the Connection Machine has lots of CPUs, and each can compute away on a piece of a problem at the same time as the others. "Massively-parallel" means a Connection Machine can have anywhere from 32 to 16,000 CPUs -- enough to search through ~~five~~ 100 years' of Wall Street Journals in a ~~second or two~~ less than a second. As Brewster noted at one

~~sex~~ sex
food
religion
death

death
religion
food
sex

Rock ~~Music~~
Drugs
self love / appearance

point, "think of a football stadium full of people, where each person takes a page or record and everybody starts searching their page when you say 'Go!'"

[VERIFY THIS FOOTNOTE.] sounds good

On the other hand, WAIS server software is available for smaller machines, such as Unix workstations. Just don't expect it to run as fast, or handle gigabytes quite as eptly.

(, as many

FOOTNOTE END

Once done, most of the WAIS clients let you save your search and even repeat it automatically, sending the you alert as new information matching the search is found (e.g., new newspaper articles detected).

Intriguingly, the server doesn't have to fully "understand" your query; natural languages can be used. To date, Kahle reports, "We have used English, French, Italian and German!"

Most WAIS servers currently available can index documents in existing formats and make them available for WAIS searching. WAIS currently supports data formats and types including text, formatted documents, pictures, spreadsheets, graphics, sound, or video, and on a variety of computer platforms.

Your WAIS client obtains a current list of available WAIS servers from the "directory of servers" list (on the Internet, maintained by TMC). spelle it out.

Each selected WAIS server reads your question and based on its words, searches the full text of the database for the most relevant documents -- i.e., documents that contain those words and phrases -- and ranks them using heuristics such as automatic word weighting.

As may be obvious, there are a few limits to this.

First, since WAIS is searching for matches of your keywords in its indexes, if the keywords aren't in the original documents, or whatever the WAIS index is based on, WAIS can't find it.

For example, if you articles that are about or refer to "tiddlywinks," you simply use keywords like "tiddlywinks". If you know any tiddlywink jargon, you may also try keywords like"squidger" and "squopp".

But you can't count on searches based on concepts, like "superhero" or "role-playing game" or "utopian fantasy" to match -- unless WAIS is searching a description of the document, and whoever added the document put the right terms in.

Second, WAIS can only search the documents it has access to, in the datasets you specify. If you search for "tiddlywinks" in the Wall Street Journal, you won't discover if there are any relevant articles in the New York Times. If an article is too new to have been added and indexed, WAIS can't find it for you.

And third, the bigger the volume of data being searched, the longer it can take. Powerful computers like TMC Connection Machines aren't cheap; many WAISed data sets, particularly those being made accessible without cost, may be run from conventional workstations and minicomputers.

How does not seem like it will be a problem, believe it or not.

Even so, WAIS adds a powerful new way to munch through those haystacks of data to find those few matching needles.

WAIS Users Report: They Like It!

The response at initial test case Peat Marwick, where WAIS was provided as a tool for $300/hour consultants. "They loved it," reports Kahle. "They were using it all them time."

As of mid-1992, developers at TMC and elsewhere in the Internet community had developed WAIS client software that can be run on a number of types of computers and operating systems, including Unix, VAX/VMS, Macintoshes, DOS, Microsoft Windows, and NeXt.

WAIS and the Internet

What does all this have to do with the Internet?

Answer: lots. The Internet began as the way WAIS developers worked together. The Internet then provided WAIS developers with a world-wide testbed of users to "take WAIS out for a spin." Today, many of the information bases about Internet resources and services, and Internet archives, have become WAIS-searchable; WAIS has also become the search tool used by people making both free-access and pay-for-view information available via the Internet.

The Internet as Developer's Tool:

Initially, as with many new computer programs, developers at organizations which were connected to the Internet used the Internet as a way to work together.

Providing a Testbed:

Started in January of 1990, WAIS developers took the next step; like Archie, WAIS client and server software, along with a number of datasets, was put onto publicly-available Internet-accessible computers. This gave WAIS developers a world-wide set of test users -- and gave the Internet community the chance to experiment with another possible solution to our continually growing flood of information whose value is only as good as our ability to find things within it.

WAIS Access to Internet Information:

By mid-1992, like Archie, Gopher, and other tools, WAIS had established itself as one of the core new tools for using the Internet, for new and experienced users alike.

Archives of many Internet mailing lists and Usenet newsgroups have been indexed under WAIS, making it *much* easier to find things. One reason: instead of having to figure out where a given archive is stored, you simply access WAIS on the Internet. The first screen is the "directory of servers," a list of what WAISed servers and databases are available, which you can do a keyword search on. Once you've identified and selected one or more WAIS servers, you then can do the search.

For example, I recently needed the electronic mail address of

someone. It was too late at night to call them, and I wanted to send them a message before I forgot.

I knew they had posted a message to the "com-priv" mailing list (a discussion about the commercialization and privitization of the Internet) at one point, and that the full archives of this mailing list were available under WAIS. So I:

1. dialed up to my account on World (a local call)

2. established a remote-login connection (using _telnet_) to the Thinking Machines' publicly-accessible WAIS system

3. Selected the "com-priv" database, by giving it as the keyword for WAIS to search in the database of servers (which is the first list offered).

4. Gave a unique string from the person's name that I knew appeared. (e.g. if you were looking for me, you might try "Daniel" and "Dern" as keywords).

5. Requested to see one of the documents with a match -- which in fact did have the person's e-mail address, as part of the message they had sent.

Elapsed time: a minute or two.

In terms of WAISed information for, about and on the Internet, as of September, 1992, TMC reports that over 300 databases have been put under Internet-accessible WAIS servers at companies and universities in over twelve countries -- and that about two new databases are registered with the Directory of Servers at TMC every week.

The "directory of WAIS servers" facility registers, lists and describes the information available on each server, including any fees for their use.

The WAIS server at Thinking Machines alone hosts over 60,000 documents including weather maps and forecasts, the CIA World Factbook, a collection of molecular biology abstracts, the Internet Info_Mac digests, and the Connection Machines Fortran manual.

These other servers include a poetry server at MIT with classical and modern poetry. The Library of Congress has plans to make their catalog available via the protocol. There's archives of e-mail discussions, documents, articles, images.... mountains of network-accessible data waiting to be searched -- from a single network point of contact.

The Internet community has given WAIS a rigorous test. Thinking Machines reports that during the past year, TMC's Internet-accessible public-access WAIS servers (including the directory of other WAIS servers) have received over 100,000 requests from users in over 200 companies and 28 countries. Plus there are about 100 downloads per day of the publicly available versions of WAIS software.

Here's what some WAIS users have to say about "what a difference a WAIS makes" (with apologies to <WHOEVER WROTE THE SONG What a
Difference a Day Makes>):

Massachusetts General Hospital (Boston, Mass), said to be the largest research hospital in the country, has been using WAIS (and Gopher) servers for a large variety of internal information since 1992.

According to Dr. J. Michael Cherry, Director of Computing for the Department of Molecular Biology at MGH, "The vast majority of computers used in our research departments are Macintoshes. The WAIS and Gopher clients are easy to set up and allow researchers to quickly start exploring our internal WAIS servers and the wealth of WAIS servers available via the Internet."

Cherry's department at MGH is providing two WAIS databases to the world, about a small flowering plant which is used as a model system within plant molecular biology. The WAIS-ed information bases contain genome database information and two years' worth of newsgroup/mailing list archives.

More use of WAIS for research and administrative information is underway at MGH. According to Cherry, "WAIS servers under development include a calendar of events, public memos, policy statements for many parts of the Hospital, and a 'Frequently Asked

Questions' database from the computer help desk." Other items under consideration include databases of information from a number of service departments within the hospital.

MIT: Student Newspaper Publication and Archives

And at MIT, where many WAIS developers studied, Reuven Lerner, System Administrator and former editor in chief of The Tech, the MIT student newspaper, reports, "We are about to launch a WAIS server for The Tech, with the last 12 years of its archives, and hope to soon begin publishing The Tech to WAIS at the same time as we print copies for distribution.

"Our WAIS server will serve both our readers (most, if not all, students and faculty at MIT are on the Internet) as well as our reporters and editors, who will use it to give better background information when assigning stories."

Practicing what they preach, developers and technical support staff and others at Thinking Machines also use WAIS on a daily basis.

"I use WAIS to index all the e-mail messages I've ever sent or received," says Kahle. "I use this as my personal memory. This application alone can change your life."

Laird Popkin in TMC's customer support division uses WAIS regularly.

"Quite frequently I need to locate a software tool to get my work done." Using WAIS, Popkin can search "all of the ReadMe files on the Internet," for example (also using Archie), use that information to determine what programs to search for, and then retrieve the software -- locating a single program from millions, across a thousand sites, in under an hour.

For example: "I receive e-mail from someone who needs to know whether we can convert "CGM" image files to something we can use.

"Since I have no idea what CGM is, I do a WAIS search of all of the ReadMe files on the Internet (courtesy of archie and ftpable-readmes.src), locating a file that contains a long list of image formats. I retrieve and read this list, so now I know what CGM

is. Then I do another WAIS search, this time of the database which has descriptions of of the files on ftp sites (ftp-list.src) and locate a Macintosh program which can read CGM files, display them, and save them as PICT files. [The same information that Archie servers are using to locate programs files -- DPD] Having found one or more sites that have this program, I transfer a copy of the file, using Xferit (since I'm working from a Macintosh running TCP/IP). I drop the CGM files in question onto the application, and everything works. Thanks to WAIS, this takes under an hour, and I didn't have to make any phone calls, send anyone e-mail, or post any queries. Without WAIS, I would probably have posted queries to one or more mailinglists and Usenet groups, not gotten responses for several days -- if at all-- and possibly then had to also query Archie."

Art Medlar <medlar@parc.xerox.com> one of the original developers of WAIS and WAIStation, currently a consultant working at Xerox in Palo Alto, says, "For me, the greatest value in WAIS is not in finding new information, but in re-finding information that I have already seen. The projects I work on generally involve several people, inconveniently spread over time and space. As a result, I use electronic mail as my primary communication medium. Prior to WAIS, one big problem with email was with finding old messages. Often, I'd find myself facing a problem which I knew to be similar to one reported and solved some months ago, but have no easy way of digging up the relevant message.

"For the past several years, I have archived every email message I have sent or received, nearly 100 megabytes worth. I keep it online and indexed. When I need to find an old message on a particular subject, I just type a few words or start a relevance-feedback search based on a similar note, and I have it within a couple of seconds. Previously, the search could have taken hours, or just been impossible.

"Other types of online documentation and information have similar problems, and WAIS provides the same sort of solution. This means it's no longer necessary to spend lots of time cataloguing and filing, and trying to remember some obscure taxonomy of files, subdirectories, and folders. Simply knowing roughly where a piece of information is, and generally what it's about, is almost always enough to find it."

BOX TEXT

WAIS Is Not Just For Internet Users

Don't make the mistake of seeing WAIS as only for Internet applications, or for use by people who have access to the Internet.

Like Archie, Gopher, TechInfo and other tools, WAIS can be run on a wide range of networked computing environments other than the Internet. "WAIS can be run over any digital network, including ISDN," says Kahle. "We've seen WAIS running in private corporate TCP/IP networks, over X.25 and even via modem connections."

A growing number of engineering, medical, government, academic and other organizations are putting WAIS to use within their own network environments. Some of these organizations are on the Internet, and make their WAISbases generally available.

WAIS systems offer as great value for corporate and university networks as it has been for the Internet (which is, in some sense, 'nothing more than' a very large enterprise network).

According to Michael L. Carroll, Manager, Advanced Computer & software Applications at Lockheed Corporation (Calabasas, CA), "At Lockheed, we have built a corporate-wide information distribution system using NetNews and WAIS. This system, called the Technology Broker System, links together several different computing environments including VM, VAX/VMS, UNIX, DOS, MS-WINDOWS and Macintohs. It makes available within Lockheed information previously inaccessible online, such as proposals, research reports, management policies and procedures, the Lockheed employee phone list, corporate library catalogs, Commerce Business Daily, and Material Safety Data Sheets (MSDS)."

According to Marc Fleischmann, in Lockheed's Information Integration
area, their WAIS servers are managing over 2.5 gigabytes of data. "The 50,000 MSDS sheets represent a gigabyte of data. Most of the other sources run from a couple of megabytes to 100 Mbs.Our server currently has 2 1.3Gb disks and they are almost full. We

expect to have 3 more 1.3 Gb disks attached to the server by the end of October.

"WAIS is a storage "magnet" and becomes disk intensive," Fleischmann acknowledges. "Data that was located on lots of different systems gets put in WAIS and the aggregate gets larger and larger. No one saw the cost of storing it on the original system but it becomes very visible when put into WAIS.".

On the other hand, he points out, "One years worth of R&D research reports (about 150 3 to 6 page reports) take up two 2" binders and weighs about 8 pounds. When put into WAIS they occupy 3 Mb of disk, about $1.50 worth of disk space at floppy prices."

Elsewhere, by now, Dow Jones & Co. should have a server available on their DowVision network that will contain several months of the Wall Street Journal and 450 business publications, and will be a for-pay server.

BOX TEXT END

Using WAIS

The following is an overall discussion of how to use WAIS. As with discussions of other Internet tools, I'll focus on the essentials; once you master these, you should be able to pursue the more advanced and version-specific features on your own.

(Also, like Archie, Gopher, Hytelnet and other Internet facilities, there are many different versions of WAIS software, for the different types of computers and access methods -- and I am quite sure there have been major new updates released and features added since this book went to press.)

Put in the explicit reminder of being on the net. not just email gateway

Get Access to WAIS:

To use WAIS, the first step is to access a WAIS client -- the end-user side of the WAIS system which will forward your queries and commands to the selected WAIS server, and display results.

Remember, WAIS is a client-server application requiring real-time Internet connections, similar to Gopher, Hytelnet, the Internet Relay Chat, and WorldWideWeb; unlike e-mail, Usenet or "ftp-by-mail" (until someone implements an e-mail interface to WAIS).

As with any Internet tool, start by seeing if a WAIS client is available on your local system, by typing "wais<RETURN>" (also try "swais<RETURN>" and "xwais<RETURN>").

(The 'swais' program is somewhat slow; if nothing happens for up to a minute after you type "swais<ENTER>" it may be initializing.

If you get an error message indicating no such command can be found, you probably don't have WAIS client software you system. *

FOOTNOTE

However, it's possible that it _is_ installed, but the setting of your PATH environment variable doesn't include the directory the WAIS command is in. If you can't this directly, contact your system administrator and ask.

FOOTNOTE END

Next, see if you can access WAIS through any of the other Internet navigation/front-end facilities currently installed in your site. Try Gopher, TechInfo, or whatever else is available. (Most Gopher servers include a gateway to the ~~Thinking Machines~~ WAIS system.)

Failing that, open a remote login connection to one of the sites offering publicly-accessible WAIS clients, such as by telnetting to quake.think.com.

Login as _wais_. You'll get SWAIS, the Simple WAIS interface for 'dumb terminals.' This is the worst of all interfaces, but:

Like most public-access guest accounts, you shouldn't need a password to access the TMC public WAIS server, but -- as the prompt will mostly likely remind you -- Internet nettiquette is to enter your Internet e-mail address here, e.g., fflint@bed.rock.org

For example:

```
world% telnet quake.think.com
Trying 192.31.181.1...
Connected to quake.think.com.
Escape character is '^]'.

SunOS UNIX (quake)

login: wais
Welcome to swais.
Please type user identifier (optional, i.e user@host):
fflint@bed.rock.com
```

SWAIS has been described by at least one Internet expert as the "least appealing of the WAIS clients, but it's an easy way to start." As of September, 1992, SWAIS lacked many of the more useful features of other WAIS client programs.

Ultimately, you'll want to get a WAIS client program installed on your machine, to take maximum advantage of your own system's user interface (e.g., Macintosh, Windows or X icons). But for now, to try WAIS out, the above are the easiest, quickest ways to go.

As you use WAIS, there may be times when the response time seems slow, or the software doesn't work perfectly. So it's important to remember that many of the programs you are using are labors of love, done by people in their spare time, and that the services on many of the computers you're accessing are being made available on a voluntary basis. (And if it works well and does what you need, it's still worth remembering the voluntary, cooperative community that made it all possible.) → send a thank you note to the maintainer.

For example, some of the WAIS clients may seem to respond slowly, or quit and break the connection unexpectedly. Trying not to "type-ahead" helps.

Here's a typical screen you might see when you login to Thinking Machines' public SWAIS: how about WAIStation?

SWAIS Source Selection Sources: 316

001: [archie.au] aarnet-resource-guide Free
001: [archie.au] aarnet-resource-guide Free
003: [weedsmunin.ub2.lu.se] academic_email_conf
Free
004: [archive.orst.edu] aeronautics Free
005: [bloat.media.mit.edu] Aesop-Fables Free
006: [nostromo.oes.orst.ed] agricultural-market-news
Free
007: [archive.orst.edu] alt.drugs Free
008: [wais.oit.unc.edu] alt.gopher Free
009: [sun-wais.oit.unc.edu] alt.sys.sun Free
010: [wais.oit.unc.edu] alt.wais Free
011: [munin.ub2.lu.se] amiga_fish_contents Free
012: [150.203.76.2] ANU-Aboriginal-Studies
$0.00/minute
013: [coombs.anu.edu.au] ANU-Asian-Religions
$0.00/minute
014: [150.203.76.2] ANU-Pacific-Linguistics
$0.00/minute
015: [coombs.anu.edu.au] ANU-Pacific-Manuscripts
Free
016: [coombs.anu.edu.au] ANU-SocSci-Netlore
$0.00/minute
017: [150.203.76.2] ANU-SSDA-Catalogues
$0.00/minute
018: [coombs.anu.edu.au] ANU-Thai-Yunnan
Free

Keywords:

<space> selects, w for keywords, arrows move, <return> searches, q quits, or ?

Select Dataset(s)/Server(s)

Once you've accessed WAIS, the next step is to select the server(s) (i.e., datasets) you want to search.

As with most Internet utilities, entering "?" or "h" gives you on-line help information.

Here's the help screen from SWAIS when you're at the "list of servers/sources" menu.

SWAIS Source Selection Help Page: 1

j, down arrow, ^N	Move Down one source
k, up arrow, ^P	Move Up one source
J, ^V, ^D	Move Down one screen
K, <esc> v, ^U	Move Up one screen
###	Position to source number ##
/sss	Search for source sss
<space>, <period>	Select current source
=	Deselect all sources
v, <comma>	View current source info
<ret>	Perform search
s	Select new sources (refresh sources list)
w	Select new keywords
X, -	Remove current source permanently
o	Set and show swais options
h, ?	Show this help display
H	Display program history
q	Leave this program

Press any key to continue

At this point, with SWAIS, you can:

o Use the arrow keys and <SPACE> to select sources

o Enter "w<RETURN>" to indicate you want to give keyword search terms for search.

Type w to enter a keyword search term, followed by <RETURN> twice to initiate the search.

[[IS IT w<RETURN> or w (pause) keyword <RET><RET>]]

Using "new" as the keyword will always find some items in each of the databases.

[[AS THE KEYWORD TO SELECTED DATABASES, OR TO ALL?]]

In response to your query, WAIS returns a list of documents that match your search criteria, listed in order of how closely they match (based on WAIS' search'n'match rules).

At this point, you can tell WAIS to retrieve a document -- display it; continue the search; or begin a new search. (Not all of the freeware WAIS client software versions may support relevance-searching at this level; i.e., you may not be able to use the "find me more like this one" search technique.)

Here's the help screen from SWAIS when you're at a list of search results:

SWAIS Search Results Help Page: 1

j, ^N	Move Down one item
k, ^P	Move Up one item
J	Move Down one screen
K	Move Up one screen
R	Show relevant documents
S	Save current item to a file
m	Mail current item to an address
##	Position to item number ##
/sss	Position to item beginning sss
<space>	Display current item
<return>	Display current item
|	Pipe current item into a unix command
v	View current item information
r	Make current item a relevant document
s	Specify new sources to search
u	Use it; add it to the list of sources
w	Make another search with new keywords
o	Set and show swais options
h	Show this help display
H	Display program history
q	Leave this program

Press any key to continue

To retrieve an item, for example, use the arrow keys and <RETURN>, or <SPACE>.

To return to the "sources" screen, enter "s".
And to quit, enter "q'"

Figures ____ show an example (provided by Thinking Machines) of WAIS being used from a Macintosh client program.

The Future of WAIS

Because WAIS can store and automatically resubmit queries, it can be the basis of your "personal window" to Internet information sources. For example, by having searches done every day or week on various newsgroups, electronic mailing lists, or other periodically updated information bases, you can get a "weekly Internet WAIS gazette" that shows you what's new in your areas of interest, where you'd otherwise have to find and sort through large volumes of new data.

Also, many other Internet services have begun using WAIS as their search engine, even though you may never see a WAIS client or build a WAIS query. Many sites' Internet Gophers, for example, send their queries to a WAIS server. The Archie servers are most likely using WAIS in some of their searching.

So when you're looking for information and you don't know where to look or how to look through it -- call on WAIS!

To Learn More About and Try WAIS

As of October, 1992, Thinking Machines was not ready to discuss anticipated "productized" versions of WAIS code (and possibly datasets too) -- i.e., code sold and supported commercially.

Meanwhile (like any self-respecting Internet application), WAIS information, code, clients and servers are freely available in a range of ways:

Freeware versions of WAIS software, specifications and documentation for WAIS clients that can run on most popular computer platforms, and for WAIS servers on several platforms, are available via public-access file transfer (anonymous-tp). Much of this comes from people who don't work at Thinking Machines, FYI.

Software developers also discuss WAIS technology via e-mail and Usenet newsgroups.

WAIS information, code, clients and servers are freely available in a range of ways:

o If you have e-mail access to the Internet, SUBSCRIBE to the mailing list:

- _wais-discussion_, a roughly-weekly digest of messages from users and developers regarding electronic publishing issues, plus info on WAIS releases (includes wais-interest messages). To subscribe, send an e-mail message to wais-discussion-request@think.com (The message should be: "add <your-email-address> wais-discussion".)

- _wais-talk_, unmoderated directly redistributed messages among developers (typically several messages per day); subscribe by sending e-mail to wais-talk-request@think.com ("add <your-email-address> wais-talk").

o If you have access the the Usenet, READ the Usenet Netnews group comp.infosystems.wais. (All postings to the wais-discussion e-mail list also go here.)

o If you have Internet anonymous-FTP access (for file transfer), using anonymous-FTP file transfer (_ftp_ to think.com_; login as user-name _anonymous_, give your Internet address in response to "Password:"), retrieve:

- freeware release of WAIS client, server and other code, including the WAIStation Macintosh program (freeware client program for accessing servers over tcp/ip) from the /wais directory,

- WAIS documentation in /pub/wais/doc/waistation_users_guide.txt

- the WAIS bibliography file /pub/wais/wais-discussion/bibliography.txt

- a Macintosh demostration screen-recorder movie put together by Steve Cisler of Apple showing some of what WAIStation does, in /wais/WAIStation-Canned-Demo.sit.hqx

- e-mail list archives on host quake.think.com in the directory /pub/wais/wais-discussion

- From WAIS client software, use WAIS servers to search through the WAIS documentation and archives.

- questions or comments about the public CM WAIS server should be directed to bug-public@think.com.

And for more information on WAIS, contact:

Barbara Lincoln Brooks
Thinking Machines Corp.
1010 El Camino Real, Suite 310 WAIS Inc
Menlo Park, CA 94025
415-329-9300
FAX: 415-329-9329
Email:barbara@think.com

[END]

UNIX Today! December 9, 1991

The Promise Of The WAIS Protocol

Emerging Standard Represents First Step Toward Unifying Data Search & Retrieval

BY JASON LEVITT

It doesn't take an expert to see that the state of modern information handling is neither open nor unified. A trip to the main library at the University of Texas at Austin—one of the top 10 college library systems in the U.S.—confirms this.

The primary card catalog is contained on an IBM mainframe accessible through various synchronous block-mode terminals scattered about the main library, and also accessible via modem through a rather crude dial-up facility.

In the main reference room, an OCLC (On-line Computer Library Center) terminal allows access to other university card catalogs; several IBM PCs are available to search CD-ROMs for bibliographical citations and abstracts on a variety of subjects; and a LEXIS/NEXUS terminal can be used for researching major U.S. court decisions. In the engineering library, an IBM PC with CD-ROM is available for searching U.S. patents.

WAIS Server Source File

```
(  :source
   :version 3
   :ip-name "nextbox.utoday.com"
   :tcp-port 5001
   :database-name "UT_TECH"
   :cost 0.00
   :cost-unit :free
   :maintainer "jason@nextbox.utoday.com"
   :description "Server created with WAIS release
   8 b2 on Mon Nov 18 16:54:19 1991 by
   jason@nextbox.utoday.com
   UNIX Today! technology articles by Jason Levitt
   The files of type text used in the index were:
   /LocalLibrary/WAIS/articles/ABCstory.txt
   /LocalLibrary/WAIS/articles/AIX3.1FS.txt
   /LocalLibrary/WAIS/articles/Benchinfo.txt
   /LocalLibrary/WAIS/articles/LPFstory.txt
   /LocalLibrary/WAIS/articles/MacXstory.txt
   /LocalLibrary/WAIS/articles/Solbourne.txt
   /LocalLibrary/WAIS/articles/SunStory.txt
   /LocalLibrary/WAIS/articles/XSerialArticle.txt
   /LocalLibrary/WAIS/articles/Xarticle.txt
   /LocalLibrary/WAIS/articles/Xcontrib.txt
)
```

Figure 1

If one were to compare information accessibility at this facility to computer resource accessibility, things here are still in the early 1980s or late '70s. Each of the systems mentioned are primarily standalone and proprietary, having their own information retrieval and organizational formats with little, if any, interoperability between databases.

While the monster mainframe card catalog might provide pointers to many sources, it is ignorant of most other on-line sources and almost never provides the most current information on subjects, despite the best efforts of its administrators. What these information-handling systems need is a dose of open systems standards and technology, the same technology that is changing the face of modern computing.

Enter WAIS, for Wide-Area Information Server, a fledgling step in the overwhelming effort needed to unify information search and retrieval technology. WAIS is an emerging open systems standard protocol for query and retrieval of information. WAIS, pronounced "ways," is the brainchild of Brewster Kahle, an employee of Thinking Machines Corp. (TMC), the No. 2 supercomputer manufacturer, behind Cray, and purveyor of fine, massively parallel systems.

The basis for WAIS is the rapidly growing electronic-publishing movement, which is seeing more and more materials, usually available only in book form, "published" or placed onto electronic media such as disk and tape, where it can be accessed with a computer.

WAIS TECHNOLOGY

WAIS is a protocol for the transmission of query and retrieval information, much like the information you would use to search a library card catalog. It is, in fact, an extension to an existing protocol standard called Z39.50, the Information Retrieval Service Definitions and Protocol Specification for Library Applications.

The Z39.50 standard was created by a group called NISO, the National Information Standards Organization, and is designed for use in electronic library card catalogs. Z39.50 essentially specifies formats for search requests directed at a database and formats for document retrieval requests. WAIS extends the Z39.50 standard to allow, among other things, discrete portions of documents, called "chunks," to be retrieved. This is especially useful in low-bandwidth situations such as serial links, where transferring an entire document in response to a query would be prohibitively time-consuming.

The WAIS protocol fits neatly at the top of the ISO 7-layer protocol model at the application and presentation layers. This makes it extremely portable to differing network environments such as TCP/IP and X.25.

Like any good open standard, the WAIS protocol does not specify or limit the technology at either end of the wire. A WAIS client can be as simple as a command line interface that takes a database name, network address and query string as input, or as complex as a combination spreadsheet and database that constantly updates in real time, based on client/server activity taking place in the background. The only condition is that the client and server exchange query and retrieval information using the WAIS protocol.

The free WAIS source code, discussed later, implements a very typical client/server model for Unix-based Internet applications. The server creates and waits on a socket attached to a well-known port. Clients attach to the port using the port number and network address of the machine. The server accepts a request, forks a child process to handle the request, and then continues to wait and service other requests.

Requests for information are largely governed by special text files maintained by the WAIS server, called "sources," that vaguely resemble library catalog cards. Figure 1 shows a source I created containing 10 of my previous technology articles for *UNIX Today!* There is enough information in the source structure, network address, TCP port number and database name for any other machine on the network running a WAIS client to locate, understand and access the information in the database.

Not surprisingly, WAIS is already being used to connect archive sites on the Internet running on various Unix-based machines as well as proprietary systems such as Macintosh and NeXT. According to Brewster Kahle, there are approximately 80 sites running public WAIS servers and many more running WAIS privately within corporations and academia. A FidoNet WAIS server site was recently added to this collection of public sites running SLIP over a 9,600-bps serial link.

On-Line WAIS Discussions And Development

alt.wais newsgroup on USENET

Join mailing lists by sending e-mail to:

wais-discussion-request@think.com – Weekly digest of mail from users and developers

wais-interest-request@think.com – Infrequent announcements of new releases

wais-talk-request@think.com – Developers' mailing list

Free WAIS client software

Clients for NeXT, X, Macintosh, Unix ASCII, GNU Emacs and Motif:
Anonymous FTP to think.com in the directory /wais

Clients for VMS, MS-DOS, Novell LAN Workplace and SunView:
Anonymous FTP to samba.oit.unc.edu in the directory /pub/wais/UNC

Free WAIS server software

Servers for NeXT and various Unix platforms
Anonymous FTP to think.com in the directory /wais

FREE WAIS SOFTWARE

I like software that you can use to get some meaningful work done quickly without having to dig too deeply into documentation. The freely available WAIS software fits that description. In the *UNIX Today!* labs, I decided to put together a small heterogeneous network and run WAIS.

General WAIS Information

Thinking Machines Corp.
1010 El Camino Real, Ste. 310
Menlo Park, CA 94025
415-329-9300 Fax: 415-329-9329

Bibliography of available WAIS documents:
Send electronic mail to: barbara@think.com

Accessing a WAIS client on the Internet:
Telnet to quake.think.com, login as: wais

Getting involved with the Nat'l Public Network:

Electronic Frontier Foundation
155 Second Street
Cambridge, MA 02141
617-864-0665
E-mail: eff@eff.org

Acting as the WAIS server system (and also a client) was a NeXTstation. Attached over Ethernet was a Macintosh running MacOS and a Sun 3/60 running SunOS 4.1. The free WAIS software included NeXT and Mac binaries and complete source code for the Unix systems, in this case the Sun. I dug out my archives of personal Unix electronic mail, about 10 Mbytes' worth, and used the indexing program included with the WAIS server to create a hashed database. I did the same with 10 of my old technology articles written for *UNIX Today!* The databases, or "sources," are listed in Figure 2.

The WAIS indexing program knows about the format of many common types of structured on-line data such as electronic mail, *netnews*, PICT-/GIF/TIFF files and biology abstract formats, and it also handles straight ASCII text.

There was also a database of WAIS documentation, created automatically by the server program, and a directory of all sources I created called "directory of information" that simply points to all the databases. After creating the databases, I ran the WAIS server program, called *waisserver*, on the NeXTstation, which sits and waits for incoming WAIS client requests.

Once the *waisserver* was running, I could access it using the clients, called WAISstations. On the Sun, which was running X/Motif, I chose to use the Motif client. I also used the Mac and NeXT WAISstations. In order to access a *waisserver*, I first had to set up my sources. Figure 3 shows a source setup window for the Mac client. I had named my database of articles "UT-TECH" on the WAIS server. The access method, "Contact," was MacTCP, Apple's TCP/IP

Continued on page 47

Focus On WAIS

Continued from page 44

protocol stack.

As shown in Figure 2, I decided to search my mail archives and technology articles for references to NCD's Xremote protocol. The results appear in the scrolling list. If the result is an entire file, such as the article contained in the file "XSerialArticle.txt," the path name for the file is listed after it.

The other results in the list are individual E-mail messages that actually are in several large text files on the WAIS server. Because the WAIS indexing program understands E-mail format, it was able to index individual E-mail messages in my E-mail archive files and transfer only those E-mail messages pertinent to the client query.

By clicking on a document in the Results window, the portion of the result most relevant to my query appears in another window. The *waisserver* uses a simplistic approach to interpreting my request for information about Xremote. It looks for the word "xremote"—the search is case-insensitive—in mail messages and headers and displays matching documents and mail messages in the results window. This turns out to be adequate as long as you put

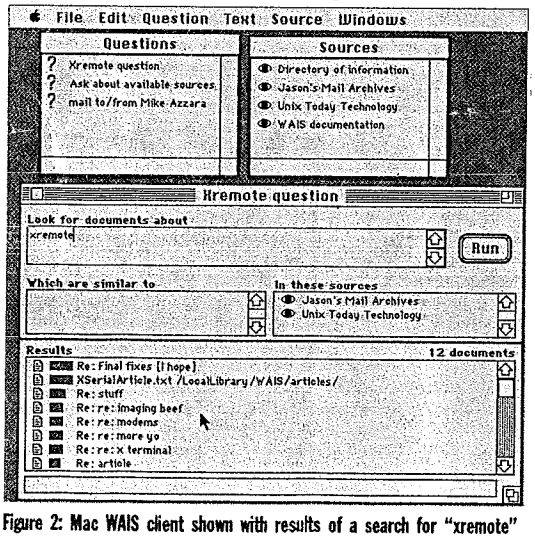

Figure 2: Mac WAIS client shown with results of a search for "xremote"

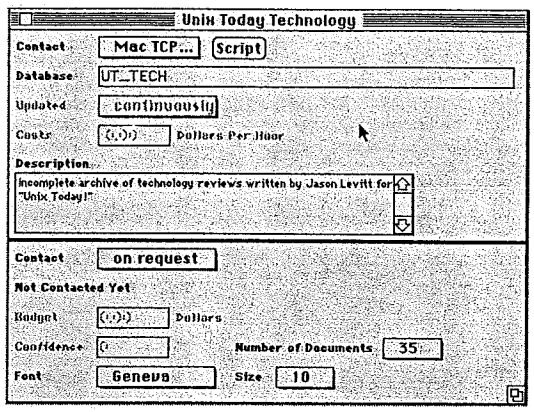

Figure 3: WAIS client set up to use Mac TCP/IP and the database UT_TECH

meaningful words in your query.

TMC has a much more sophisticated searching mechanism in its Internet server, *quake.think.com*; however, the search source code is not freely available.

One of the key features of the WAIS protocol is its ability to allow secondary search criteria. In Figure 2, the criteria would be entered by copying a result, or chunk of a result, to the "which are similar to" window. A subsequent search would use any words contained in that window as additional search criteria. Repeatedly using that method can quickly refine the search parameters.

AN OPEN END

The next version of the WAIS protocol should be officially folded into the Z39.50 standard this month and is expected to include multimedia support and integral support for English-language queries. These enhancements should add considerable clout to WAIS, given the infant state of commercial multimedia query/retrieval technology.

WAIS software is freely available from a number of sites. Unfortunately, the WAIS client program can only be obtained via anonymous FTP at this time, which means you have to have direct Internet access.

The WAIS server and X-based client program for Unix are available on *uunet.uu.net* in the directory */networking/distrib-is/wais*.

My small network experiment with WAIS only touched on its full potential; however, for my small database needs, it was quite useful. The free WAIS software is, like the MIT X software, meant as refer-

Continued on page 48

Spotlight On WAIS

Continued from page 47

ence software for further development, not as a commercial-quality implementation.

I encountered bugs, such as a persistent permissions error from the NeXT client, and strange window clipping from the Motif client, and I have yet to get the *waisserver* running cleanly under SVR4. But when the software is open and free, who cares?

The vision of WAIS is not only easy access, retrieval and publishing of information, but the creation of a marketplace that can encourage new information sources.

That, according to advocacy groups such as the Electronic Frontier Foundation, could be realized through ISDN, an infrastructure for a "National Public Network" that already is partially implemented in the U.S. telephone system. Such a network could bring the reality of WAIS-based online information services into virtually every home.